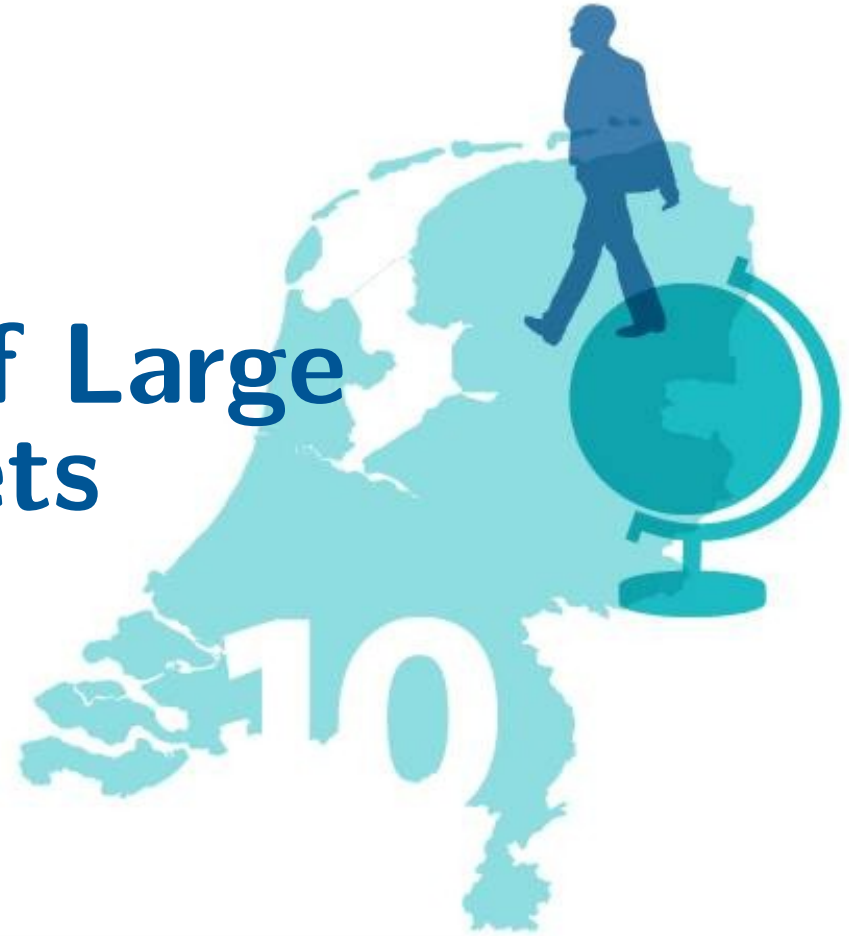
Visual Profiling of Large Statistical Datasets



Martijn Tennekes
Edwin de Jonge, Piet Daas

January 31, 2011

BLUE ETS
EUROPEAN COMMISSION
European Research Area
SEVENTH FRAMEWORK PROGRAMME
Funded under Socio-economic Sciences & Humanities

Centraal Bureau voor de Statistiek

Outline

- Introduction
- Tableplot description
- Applications
- Implementation in R

Introduction

Large statistical dataset

- Administrative sources
- Survey data

Quality assessment at a technical level

Step 1: Technical checks (e.g. readability and convertability)

Step 2: Data profiling

- Representation and distribution of values
- Strange data patterns
- Occurrence of missing values

Introduction

Large statistical dataset

- Administrative sources
- Survey data

Quality assessment at a technical level

Step 1: Technical checks (e.g. readability and convertability)

Step 2: Data profiling

- Representation and distribution of values
- Strange data patterns
- Occurrence of missing values

Introduction

Large statistical dataset

- Administrative sources
- Survey data

Quality assessment at a technical level

Step 1: Technical checks (e.g. readability and convertability)

Step 2: Data profiling

- Representation and distribution of values
- Strange data patterns
- Occurrence of missing values

Introduction

Large statistical dataset

- Administrative sources
- Survey data

Quality assessment at a technical level

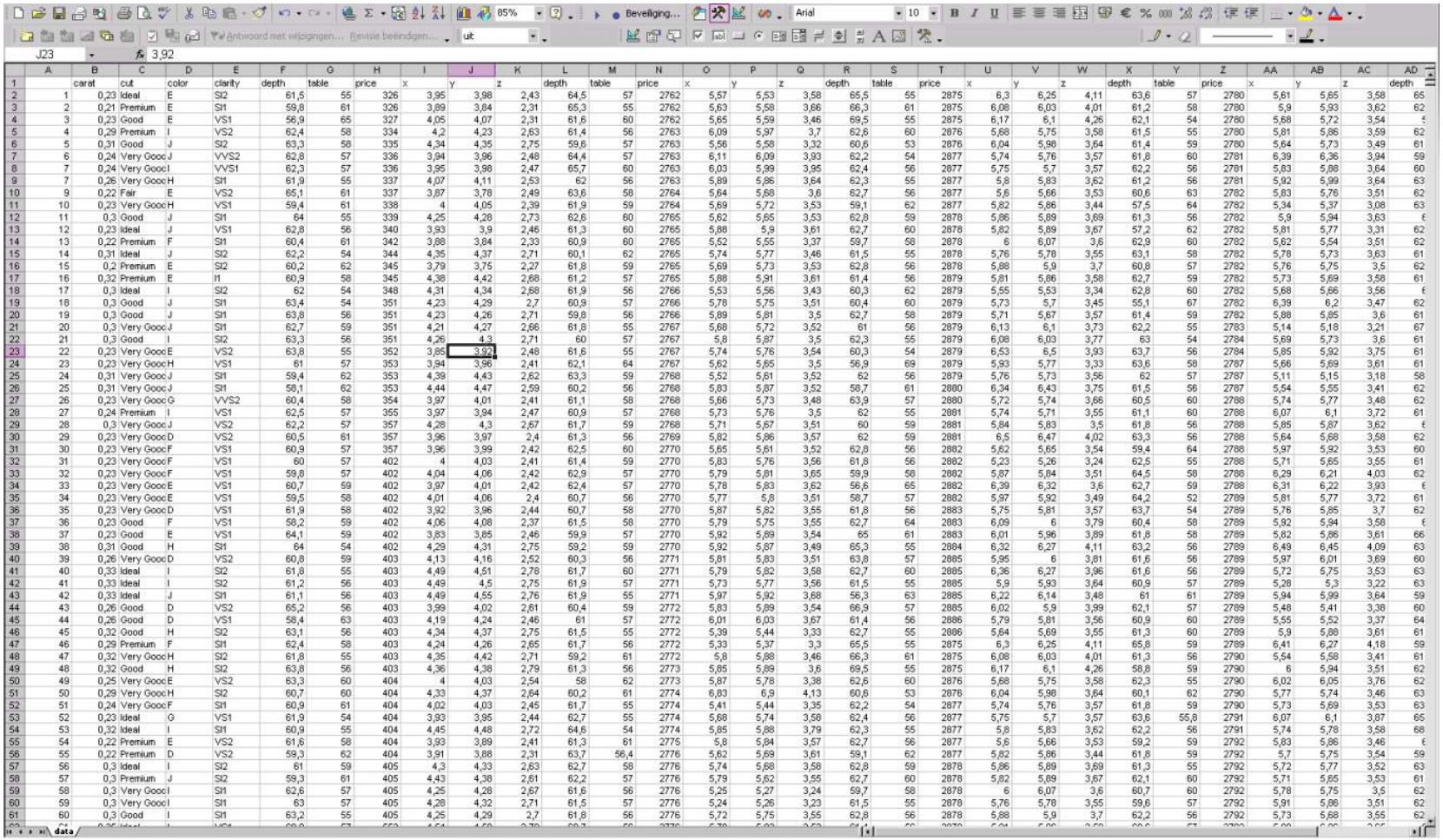
Step 1: Technical checks (e.g. readability and convertability)

Step 2: Data profiling

- Representation and distribution of values
- Strange data patterns
- Occurrence of missing values

Introduction

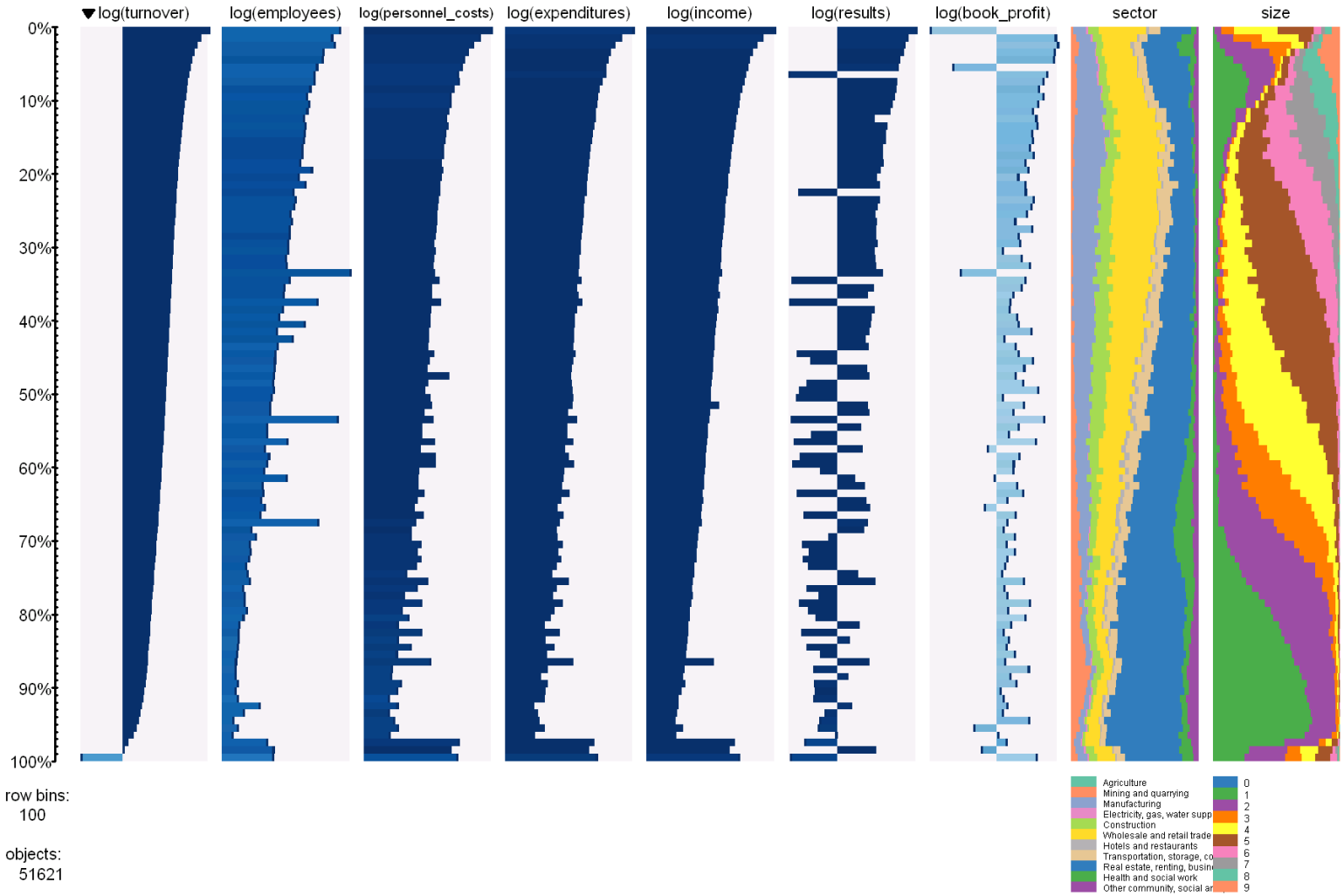
Traditional approach:



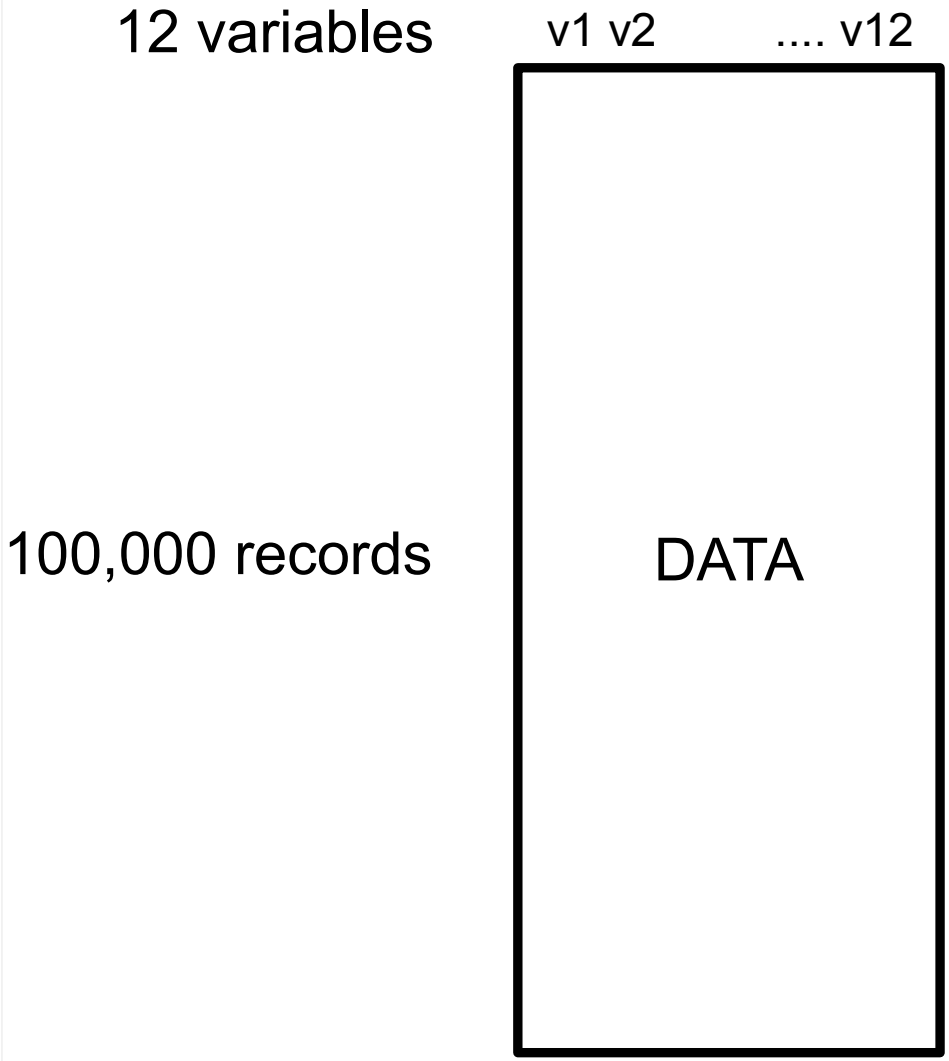
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1		carat	cut	color	clarity	depth	table	price	x	y	z	depth	table	price	x	y	z	depth	table	price	x	y	z	depth	table	price	x	y	z	depth
2	1	0,23	Ideal	E	SI2	61,5	55	326	3,95	3,98	2,43	64,5	57	2762	5,57	5,53	3,58	65,5	55	2875	6,3	6,25	4,11	63,6	57	2780	5,61	5,65	3,58	65
3	2	0,21	Premium	E	SI1	59,8	61	326	3,89	3,84	2,31	65,3	55	2762	5,63	5,58	3,66	66,3	61	2875	6,08	6,03	4,01	61,2	58	2780	5,9	5,93	3,62	62
4	3	0,23	Good	E	VS1	56,9	65	327	4,05	4,07	2,31	61,6	60	2762	5,65	5,59	3,46	69,5	55	2875	6,17	6,1	4,26	62,1	54	2780	5,68	5,72	3,54	
5	4	0,29	Premium	I	VS2	62,4	58	334	4,2	4,23	2,63	61,4	56	2763	6,09	5,97	3,7	62,6	60	2876	5,68	5,75	3,58	61,5	55	2780	5,81	5,86	3,59	62
6	5	0,31	Good	J	SI2	63,3	58	335	4,34	4,35	2,75	59,6	57	2763	5,56	5,58	3,32	60,6	53	2876	6,04	5,98	3,64	61,4	59	2780	5,64	5,73	3,49	61
7	6	0,24	Very Good	J	VVS2	62,8	57	336	3,94	3,96	2,48	64,4	57	2763	6,11	6,09	3,93	62,2	54	2877	5,74	5,76	3,57	61,8	60	2781	6,39	6,36	3,94	59
8	7	0,24	Very Good	I	VVS1	62,3	57	336	3,95	3,98	2,47	65,7	60	2763	6,03	5,99	3,95	62,4	56	2877	5,75	5,7	3,57	62,2	56	2781	5,83	5,88	3,64	60
9	7	0,26	Very Good	H	SI1	61,9	55	337	4,07	4,11	2,53	62	56	2763	5,89	5,86	3,64	62,3	55	2877	5,8	5,83	3,62	61,2	56	2781	5,92	5,99	3,64	63
10	9	0,22	Fair	E	VS2	65,1	61	337	3,87	3,78	2,49	63,6	58	2764	5,64	5,68	3,6	62,7	56	2877	5,6	5,66	3,53	60,6	63	2782	5,83	5,76	3,51	62
11	10	0,23	Very Good	H	VS1	59,4	61	338	4	4,05	2,39	61,9	59	2764	5,69	5,72	3,53	59,1	62	2877	5,82	5,86	3,44	57,5	64	2782	5,34	5,37	3,08	63
12	11	0,3	Good	J	SI1	64	55	339	4,25	4,28	2,73	62,6	60	2765	5,62	5,65	3,53	62,8	59	2878	5,86	5,89	3,69	61,3	56	2782	5,9	5,94	3,63	
13	12	0,23	Ideal	J	VS1	62,8	56	340	3,93	3,9	2,46	61,3	60	2765	5,88	5,9	3,61	62,7	60	2878	5,82	5,89	3,67	57,2	62	2782	5,81	5,77	3,31	62
14	13	0,22	Premium	F	SI1	60,4	61	342	3,88	3,84	2,33	60,9	60	2765	5,52	5,55	3,37	59,7	58	2878	6	6,07	3,6	62,9	60	2782	5,62	5,54	3,51	62
15	14	0,31	Ideal	J	SI2	62,2	54	344	4,35	4,37	2,71	60,1	62	2765	5,74	5,77	3,46	61,5	55	2878	5,76	5,78	3,55	63,1	58	2782	5,78	5,73	3,63	61
16	15	0,2	Premium	E	SI2	60,2	62	345	3,79	3,75	2,27	61,8	59	2765	5,69	5,73	3,53	62,8	56	2878	5,88	5,9	3,7	60,8	57	2782	5,76	5,75	3,5	62
17	16	0,32	Premium	E	I1	60,9	58	345	4,38	4,42	2,68	61,2	57	2765	5,88	5,91	3,61	61,4	56	2879	5,81	5,86	3,58	62,7	59	2782	5,73	5,69	3,58	61
18	17	0,3	Ideal	I	SI2	62	54	348	4,31	4,34	2,68	61,9	56	2766	5,53	5,56	3,43	60,3	62	2879	5,55	5,53	3,34	62,8	60	2782	5,68	5,66	3,56	
19	18	0,3	Good	J	SI1	63,4	54	351	4,23	4,29	2,7	60,9	57	2766	5,78	5,75	3,51	60,4	60	2879	5,73	5,7	3,45	55,1	67	2782	6,39	6,2	3,47	62
20	19	0,3	Good	J	SI1	63,8	56	351	4,23	4,26	2,71	59,8	56	2766	5,89	5,81	3,5	62,7	58	2879	5,71	5,67	3,57	61,4	59	2782	5,88	5,85	3,6	61
21	20	0,3	Very Good	J	SI1	62,7	59	351	4,21	4,27	2,66	61,8	55	2767	5,68	5,72	3,52	61	56	2879	6,13	6,1	3,73	62,2	55	2783	5,14	5,18	3,21	67
22	21	0,3	Good	I	SI2	63,3	56	351	4,26	4,3	2,71	60	57	2767	5,8	5,87	3,5	62,3	55	2879	6,08	6,03	3,77	63	54	2784	5,69	5,73	3,6	61
23	22	0,23	Very Good	E	VS2	63,8	55	352	3,85	3,92	2,48	61,6	55	2767	5,74	5,76	3,54	60,3	54	2879	6,53	6,5	3,93	63,7	56	2784	5,85	5,92	3,75	61
24	23	0,23	Very Good	H	VS1	61	57	353	3,94	3,96	2,41	62,1	64	2767	5,62	5,65	3,5	56,9	69	2879	5,93	5,77	3,33	63,6	58	2787	5,66	5,69	3,61	61
25	24	0,31	Very Good	J	SI1	59,4	62	353	4,39	4,43	2,62	63,3	59	2768	5,52	5,61	3,52	62	56	2879	5,76	5,73	3,56	62	57	2787	5,11	5,15	3,18	58
26	25	0,31	Very Good	J	SI1	58,1	62	353	4,44	4,47	2,59	60,2	56	2768	5,83	5,87	3,52	58,7	61	2880	6,34	6,43	3,75	61,5	56	2787	5,54	5,55	3,41	62
27	26	0,23	Very Good	G	VVS2	60,4	58	354	3,97	4,01	2,41	61,1	58	2768	5,66	5,73	3,48	63,9	57	2880	5,72	5,74	3,66	60,5	60	2788	5,74	5,77	3,48	62
28	27	0,24	Premium	I	VS1	62,5	57	355	3,97	3,94	2,47	60,9	57	2768	5,73	5,76	3,5	62	55	2881	5,74	5,71	3,55	61,1	60	2788	6,07	6,1	3,72	61
29	28	0,3	Very Good	J	VS2	62,2	57	357	4,28	4,3	2,67	61,7	59	2768	5,71	5,67	3,51	60	59	2881	5,84	5,83	3,5	61,8	56	2788	5,85	5,87	3,62	
30	29	0,23	Very Good	D	VS2	60,5	61	357	3,96	3,97	2,4	61,3	56	2769	5,82	5,86	3,57	62	59	2881	6,5	6,47	4,02	63,3	56	2788	5,64	5,68	3,58	62
31	30	0,23	Very Good	F	VS1	60,9	57	357	3,96	3,99	2,42	62,5	60	2770	5,65	5,61	3,52	62,8	56	2882	5,62	5,65	3,54	59,4	64	2788	5,97	5,92	3,53	60
32	31	0,23	Very Good	F	VS1	60	57	402	4	4,03	2,41	61,4	59	2770	5,83	5,76	3,56	61,8	56	2882	5,23	5,26	3,24	62,5	55	2788	5,71	5,65	3,55	61
33	32	0,23	Very Good	F	VS1	59,8	57	402	4,04	4,06	2,42	62,9	57	2770	5,79	5,81	3,65	59,9	58	2882	5,87	5,84	3,51	64,5	58	2788	6,29	6,21	4,03	62
34	33	0,23	Very Good	E	VS1	60,7	59	402	3,97	4,01	2,42	62,4	57	2770	5,78	5,83	3,62	56,6	65	2882	6,39	6,32	3,6	62,7	59	2788	6,31	6,22	3,93	
35	34	0,23	Very Good	E	VS1	59,5	58	402	4,01	4,06	2,4	60,7	56	2770	5,77	5,8	3,51	58,7	57	2882	5,97	5,92	3,49	64,2	52	2789	5,81	5,77	3,72	61
36	35	0,23	Very Good	D	VS1	61,9	58	402	3,92	3,96	2,44	60,7	58	2770	5,87	5,82	3,55	61,8	56	2883	5,75	5,81	3,57	63,7	54	2789	5,76	5,85	3,7	62
37	36	0,23	Good	F	VS1	58,2	59	402	4,06	4,08	2,37	61,5	58	2770	5,79	5,75	3,55	62,7	64	2883	6,09	6	3,79	60,4	58	2789	5,92	5,94	3,58	
38	37	0,23	Good	E	VS1	64,1	59	402	3,83	3,85	2,46	59,9	57	2770	5,92	5,89	3,54	65	61	2883	6,01	5,96	3,89	61,8	58	2789	5,82	5,86	3,61	66
39	38	0,31	Good	H	SI1	64	54	402	4,29	4,31	2,75	59,2	59	2770	5,92	5,87	3,49	65,3	55	2884	6,32	6,27	4,11	63,2	56	2789	6,49	6,45	4,09	63
40	39	0,26	Very Good	D	VS2	60,8	59	403	4,13	4,16	2,52	60,3	56	2771	5,81	5,83	3,51	63,8	57	2885	5,95	6	3,81	61,6	56	2789	5,97	6,01	3,69	60
41	40	0,33	Ideal	I	SI2	61,8	55	403	4,49	4,51	2,78	61,7	60	2771	5,79	5,82	3,58	62,7	60	2885	6,36	6,27	3,96	61,6	56	2789	5,72	5,75	3,53	63
42	41	0,33	Ideal	I	SI2	61,2	56	403	4,49	4,5	2,75	61,9	57	2771	5,73	5,77	3,56	61,5	55	2885	5,9	5,93	3,64	60,9	57	2789	5,28	5,3	3,22	63
43	42	0,33	Ideal	J	SI1	61,1	56	403	4,49	4,55	2,76	61,9	55	2771	5,97	5,92	3,68	56,3	63	2885	6,22	6,14	3,48	61	61	2789	5,94	5,99	3,64	59
44	43	0,26	Good	D	VS2	65,2	56	403	3,99	4,02	2,61	60,4	59	2772	5,83	5,89	3,54	66,9	57	2885	6,02	5,9	3,99	62,1	57	2789	5,48	5,41	3,38	60
45	44	0,26	Good	D	VS1	58,4	63	403	4,19	4,24	2,46	61	57	2772	6,01	6,03	3,67	61,4	56	2886	5,79	5,81	3,56	60,9	60	2789	5,55	5,52	3,37	64
46	45	0,32	Good	H	SI2	63,1	56	403	4,34	4,37	2,75	61,5	55	2772	5,39	5,44	3,33	62,7	55	2886	5,64	5,69	3,55	61,3	60	2789	5,9	5,88	3,61	61
47	46	0,29	Premium	F	SI1	62,4	58	403	4,24	4,26	2,65	61,7	56	2772	5,33	5,37	3,3	65,5	55	2875	6,3	6,25	4,11	65,8	59	2789	6,41	6,27	4,18	59
48	47	0,32	Very Good	H	SI2	61,8	55	403	4,35	4,42	2,71	59,2	61	2772	5,8	5,88	3,46	66,3	61	2875	6,08	6,03	4,01	61,3	56	2790	5,54	5,58	3,41	61
49	48	0,32	Good	H	SI2	63,8	56	403	4,36	4,38	2,79	61,3	56	2773	5,85	5,89	3,6	69,5	55	2875	6,17	6,1	4,26	58,8	59	2790	6	5,94	3,51	62
50	49	0,25	Very Good	E	VS2	63,3	60	404	4	4,03	2,54	58	62	2773	5,87	5,78	3,38	62,6	60	2876	5,68	5,75	3,58	62,3	55	2790	6,02	6,05	3,76	62
51	50	0,29	Very Good	H	SI2	60,7	60	404	4,33	4,37	2,64	60,2	61	2774	6,83	6,9	4,13	60,6	53	2876	6,04	5,98	3,64	60,1	62	2790	5,77	5,74	3,46	63
52	51	0,24	Very Good	F	SI1	60,9	61	404	4,02	4,03	2,45	61,7	55	2774	5,41	5,44	3,35	62,2	54	2877	5,74	5,76	3,57	61,8	59	2790	5,73	5,69	3,53	63
53	52	0,23	Ideal	G	VS1	61,9	54	404	3,93	3,95	2,44	62,7	55	2774	5,68	5,74	3,58	62,4	56	2877	5,75	5,7	3,57	63,6	55,8	2791	6,07	6,1	3,87	65
54	53	0,32	Ideal	I	SI1	60,9	55	404	4,45	4,48	2,72	64,6	54	2774	5,85	5,88	3,79	62,3	55	2877	5,8	5,83	3,62	62,2	56	2791	5,74	5,78	3,58	68
55	54	0,22	Premium	E	VS2	61,6	58	404	3,93	3,89	2,41	61,3	61	2775	5,8	5,84	3,57	62,7	56	2877	5,6	5,66	3,53	59,2	59	2792	5,83	5,86	3,46	
56	55	0,22	Premium	D	VS2	59,3	62	404	3,91	3,88	2,31	63,7	56,4	2776	5,62	5,69	3,61	59,1	62	2877	5,82	5,86	3,44	61,8	59	2792	5,7	5,75	3,54	59
57	56	0,3	Ideal	I	SI2	61	59	405	4,3	4,33	2,63	62,7	58	2776	5,74	5,68	3,58	62,8	59	2878	5,86	5,89	3,69	61,3	55	2792	5,72	5,77	3,52	63
58	57	0,3	Premium	J	SI2	59,3	61	405	4,43	4,38	2,61	62,2	57	2776	5,79	5,62	3,55	62,7	60	2878	5,82	5,89	3,67	62,1	60	2792	5,71	5,65	3,53	61
59	58	0,3	Very Good	I	SI1	62,6	57	405	4,25	4,28	2,67	61,6	56	2776	5,25	5,27	3,24	59,7	58	2878	6	6,07	3,6	60,7	60	2792	5,78	5,75	3,5	62
60	59	0,3	Very Good	I	SI1	63	57	405	4,28	4,32	2,71	61,5	57	2776	5,24	5,26	3,23	61,5	55	2878	5,76	5,78	3,55	59,6	57	2792	5,91	5,86	3,51	62
61	60	0,3	Good	I	SI1	63,2	55	405	4,25	4,29	2,7	61,8	56	2776	5,72	5,75	3,55	62,8	56	2878	5,88	5,9	3,7	62,2	56	2792	5,73	5,68	3,55	62

Introduction

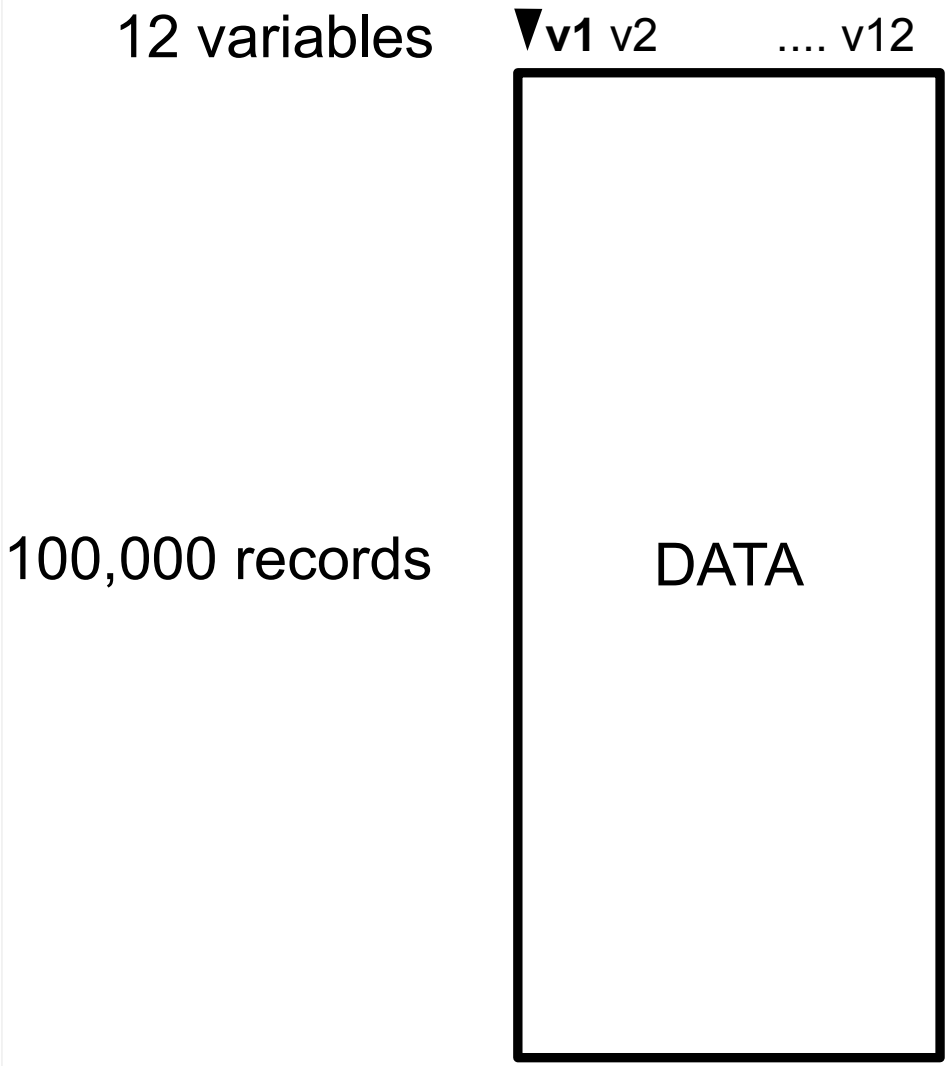
New approach:



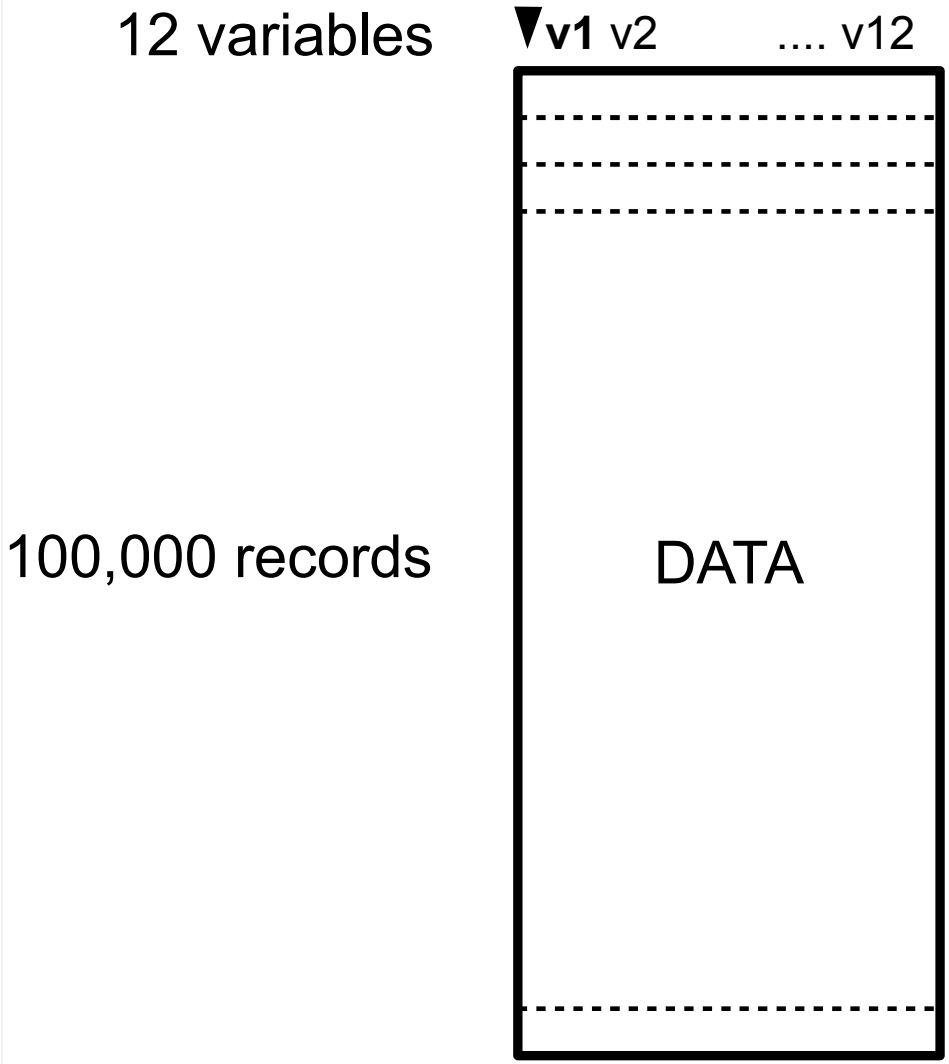
Tableplot description



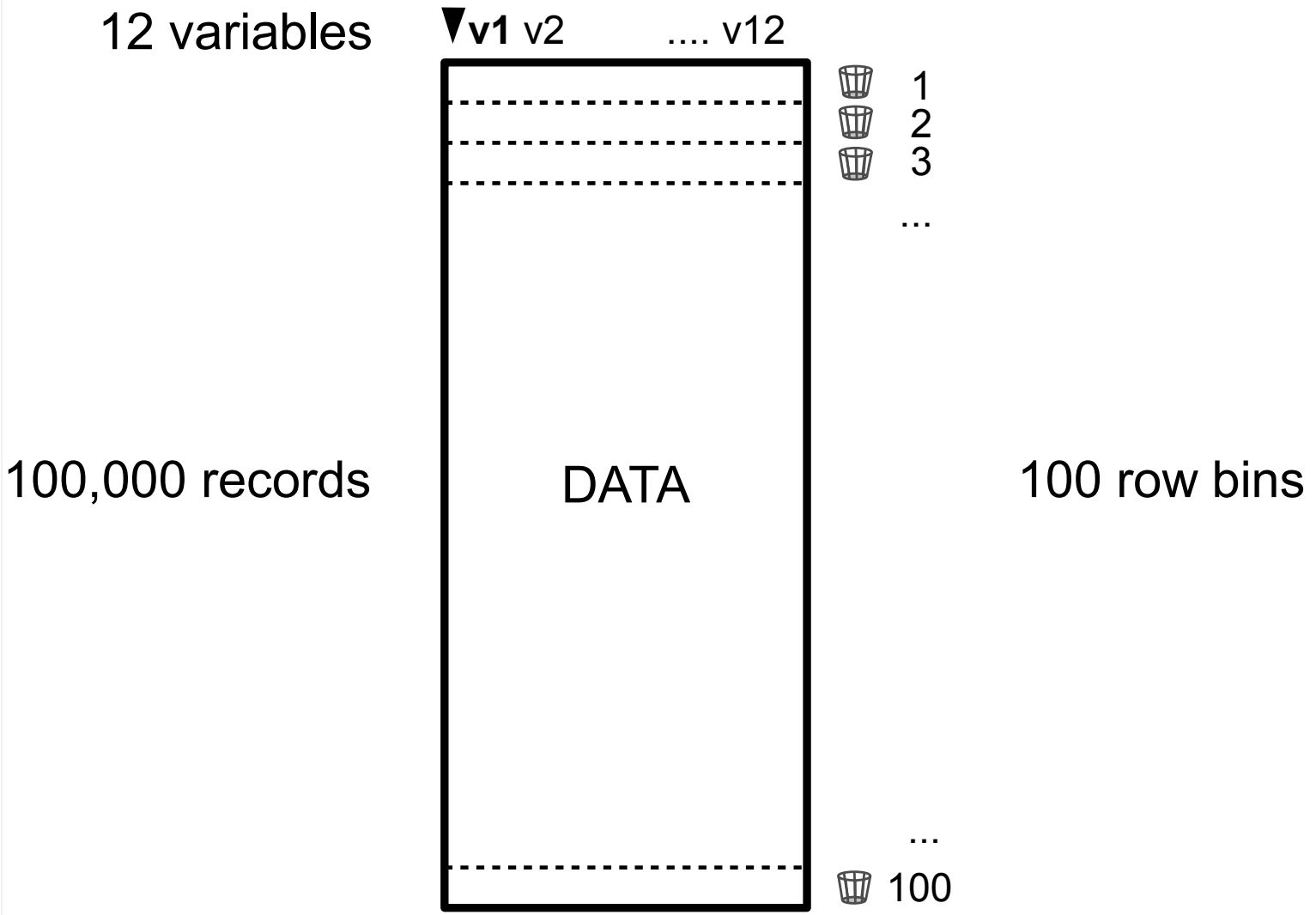
Tableplot description



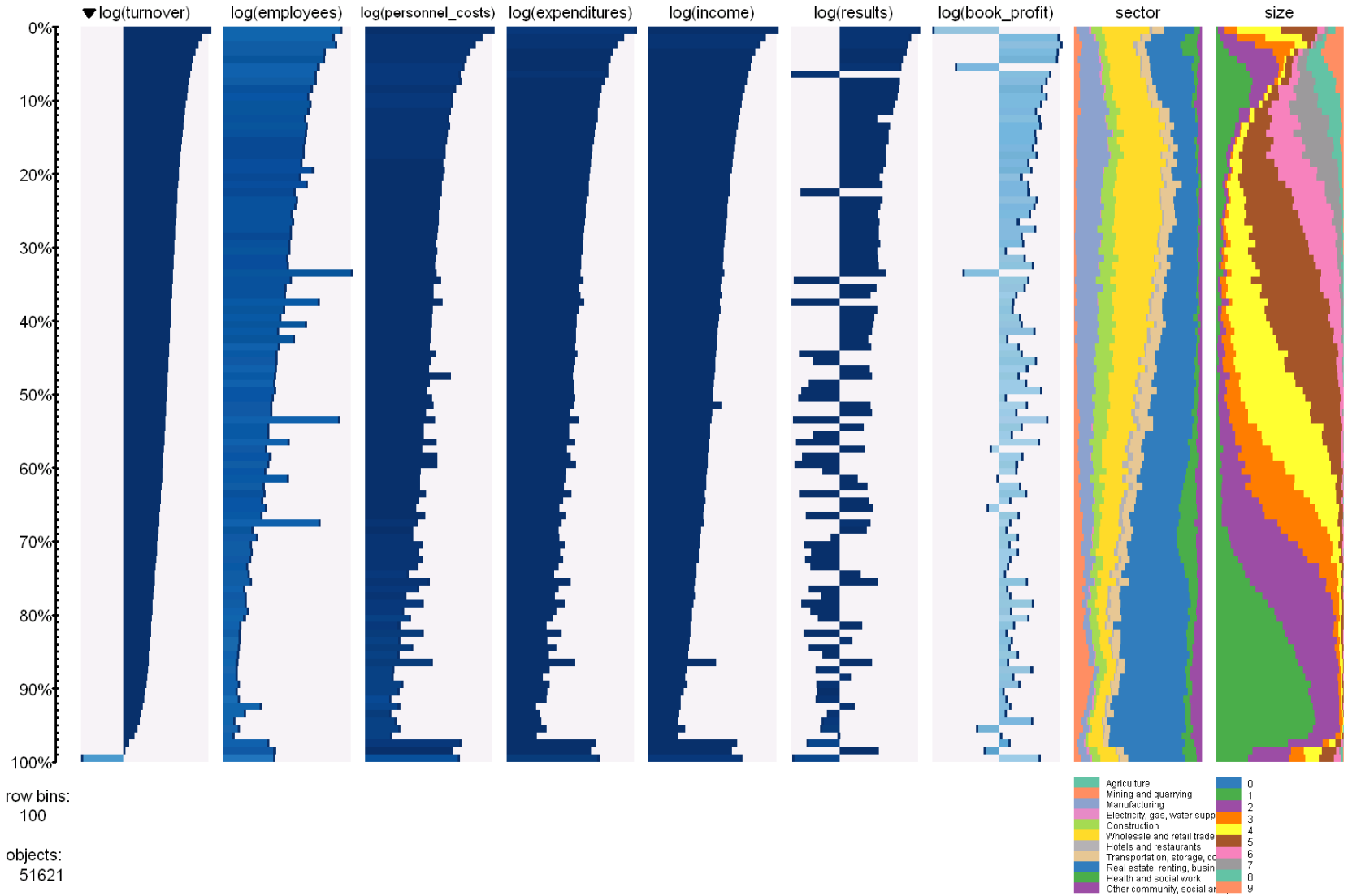
Tableplot description



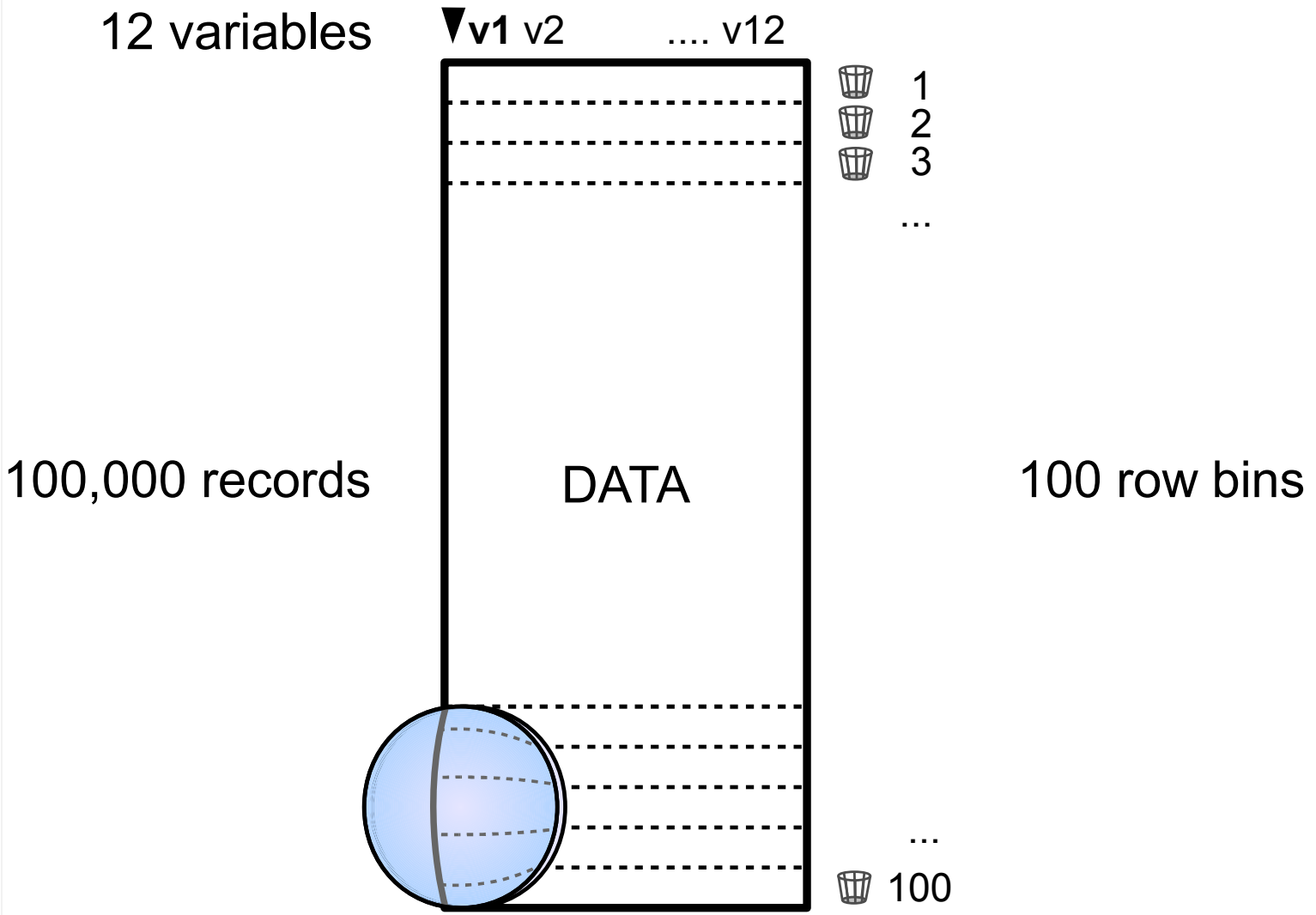
Tableplot description



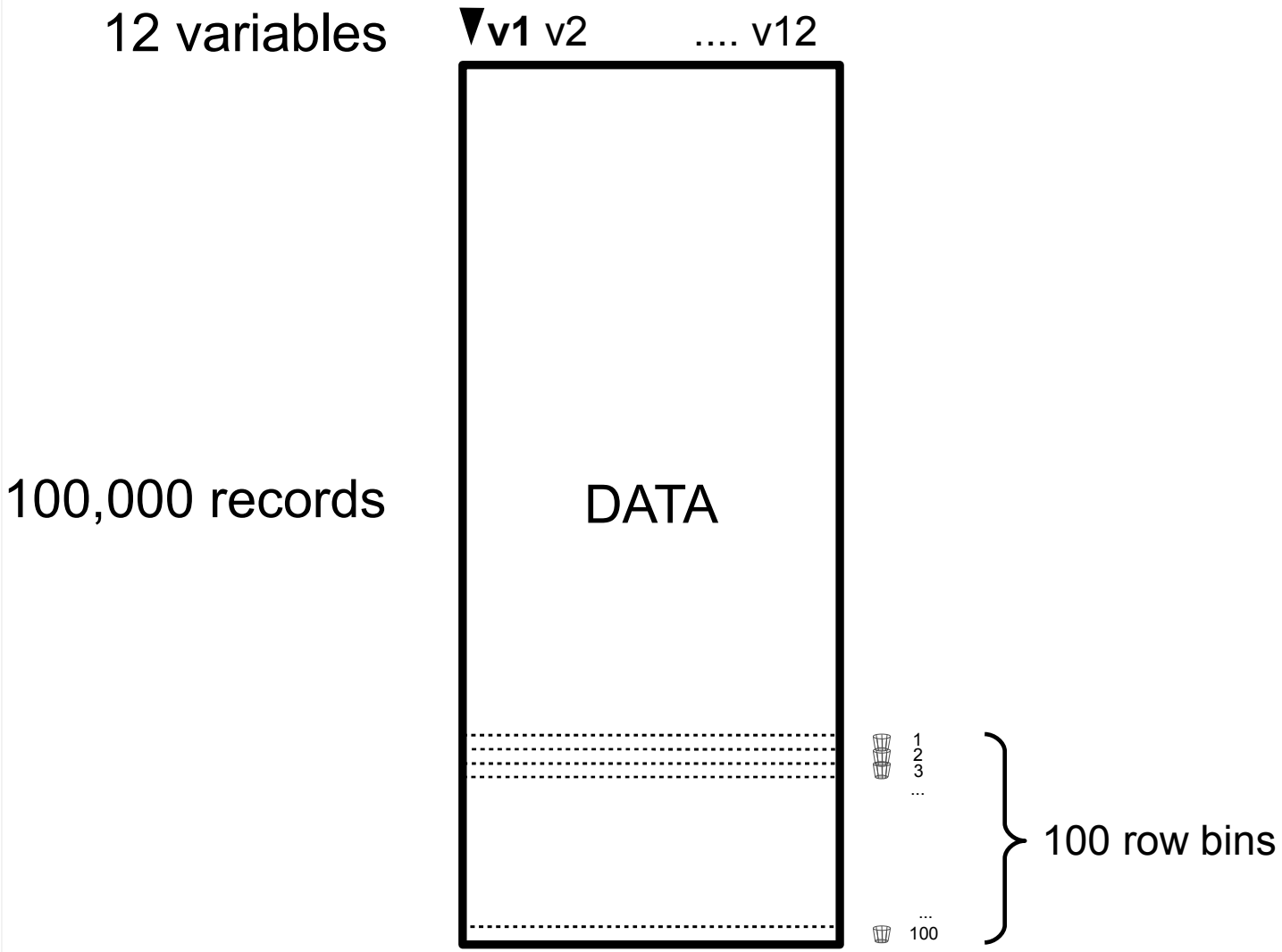
Tableplot description



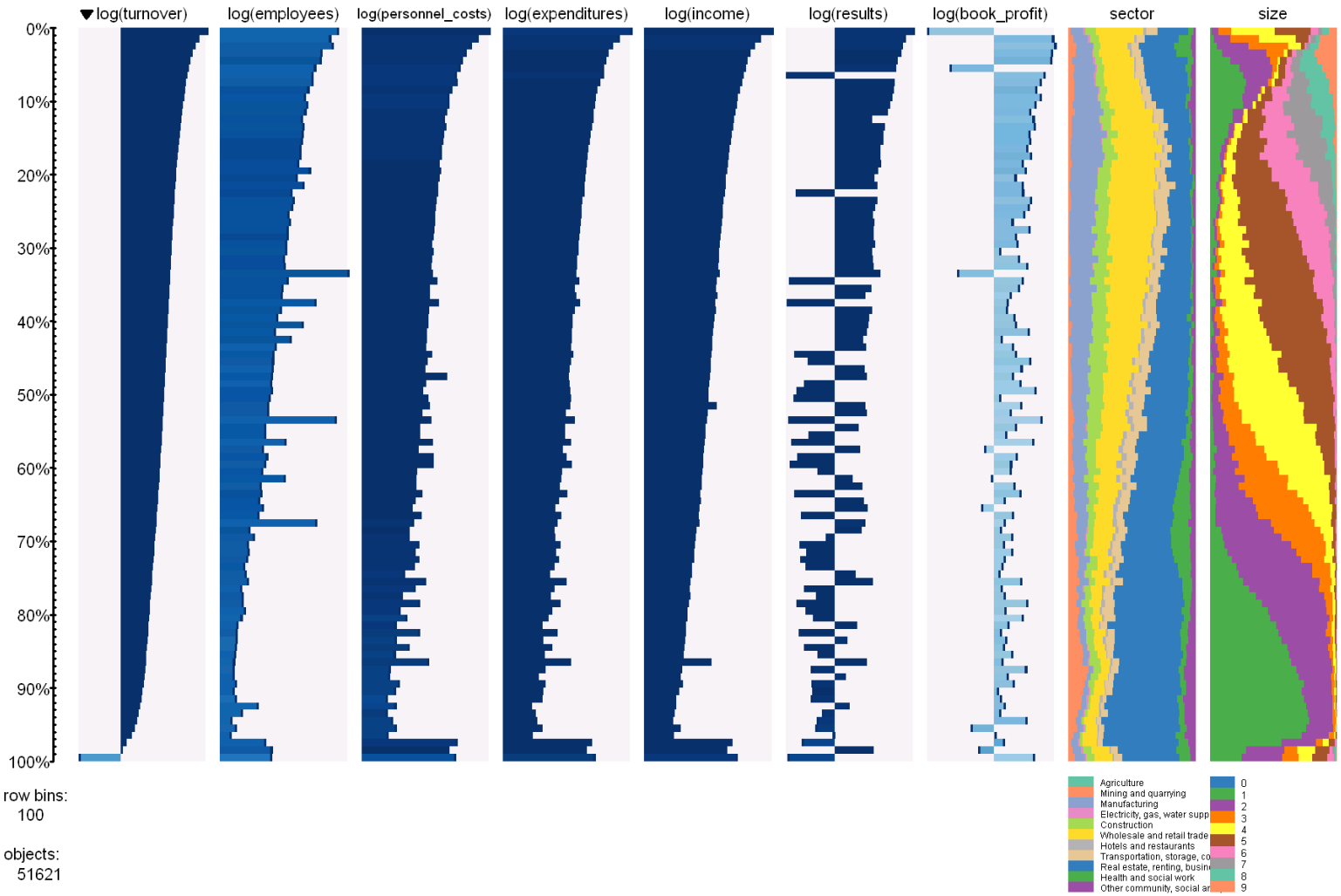
Tableplot description



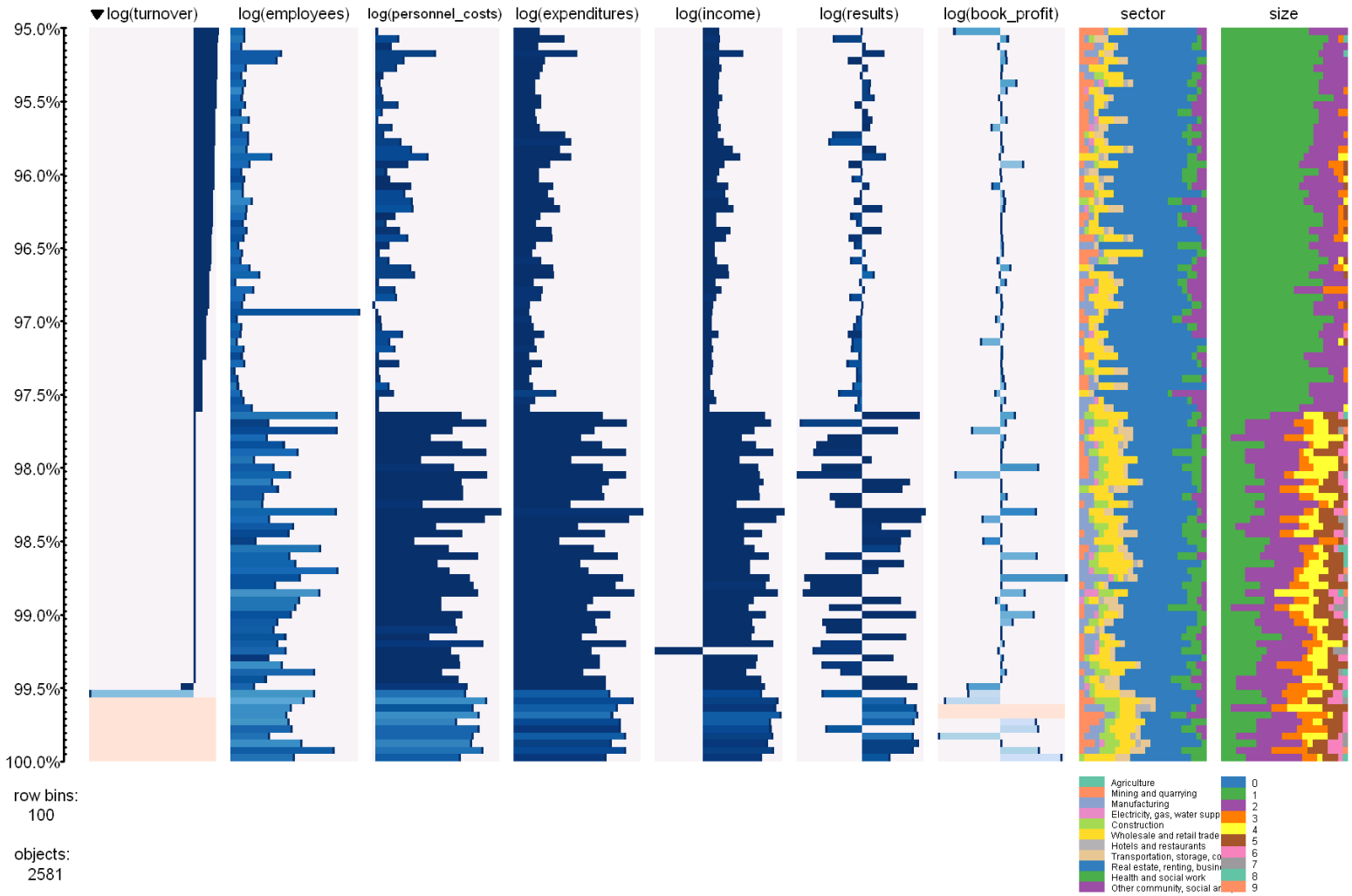
Tableplot description



Tableplot description



Tableplot description



Tableplot description

Quality measures:

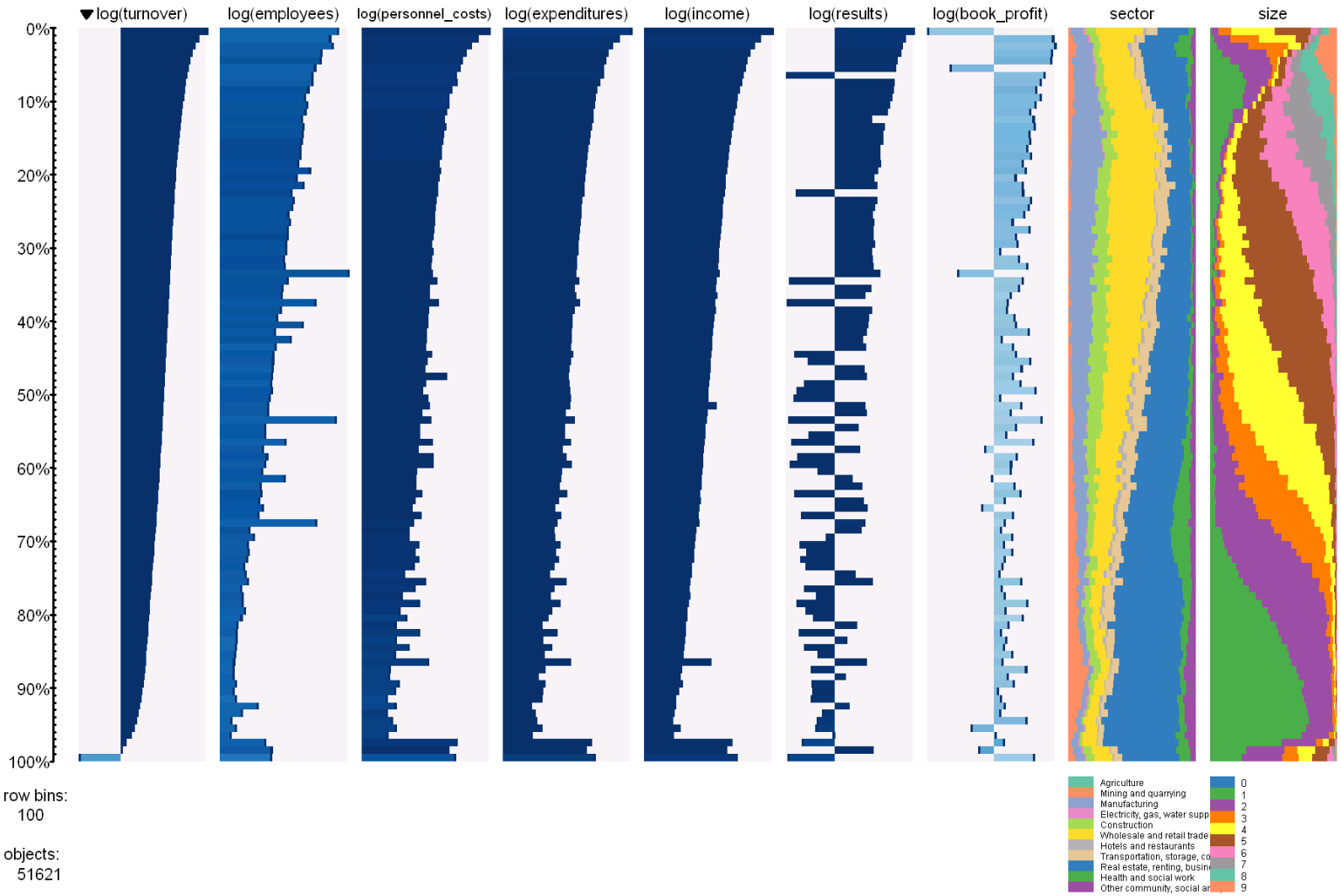
1. Smoothness of a data distribution
2. Selectivity of missing values
3. Distribution of correlated variables

Applications

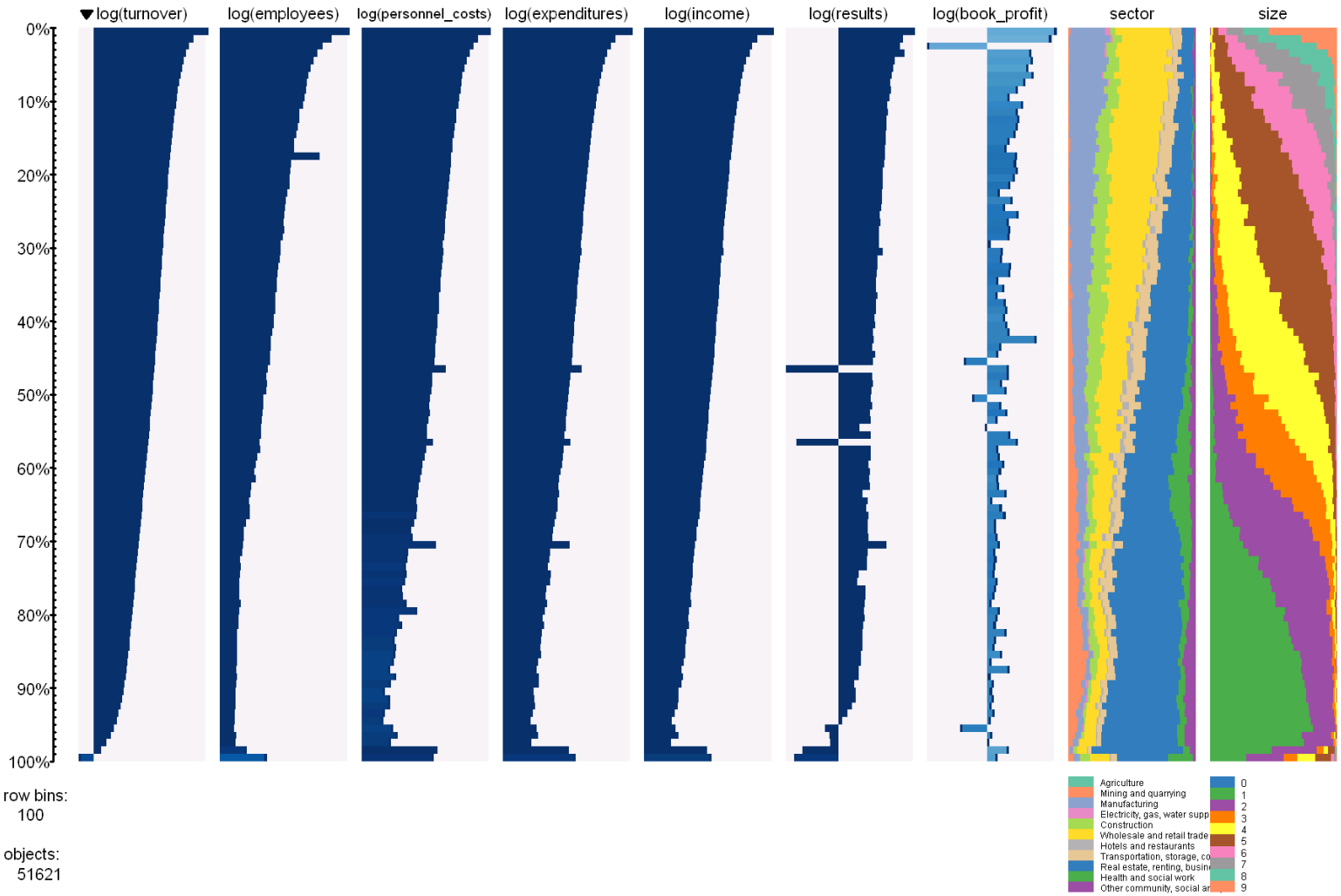
Structural Business Statistics (SBS)

- Large business survey
- Circa 50,000 respondents
- Data editing and analysis process:
  1. Unprocessed data
  2. Edited data
  3. Data prepared for analysis

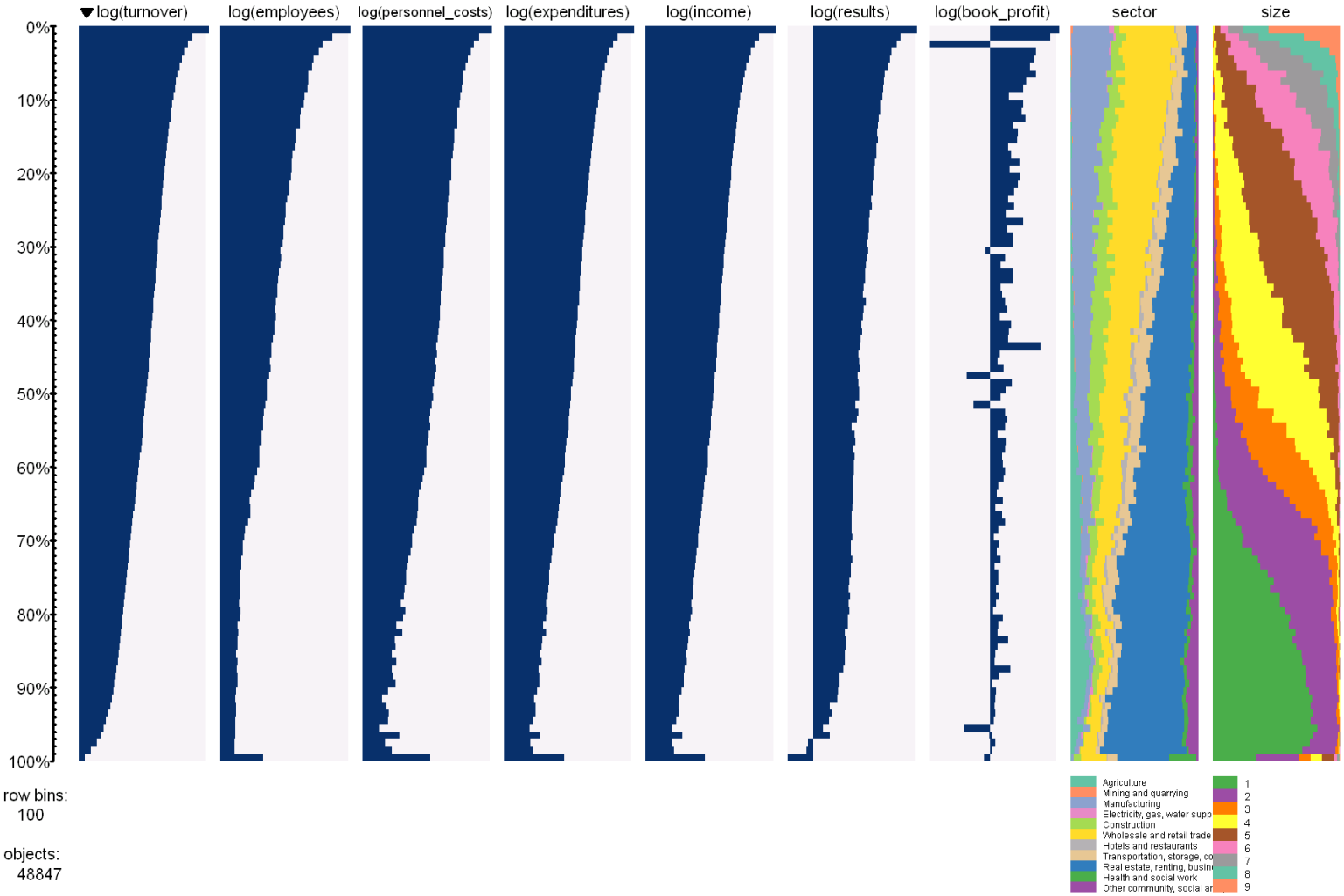
Unprocessed data



Edited data


log(turnover)
log(employees)
log(personnel_costs)
log(expenditures)
log(income)
log(results)
log(book_profit)
sector
size
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
row bins:
100
objects:
51621
Agriculture
Mining and quarrying
Manufacturing
Electricity, gas, water supp
Construction
Wholesale and retail trade
Hotels and restaurants
Transportation, storage, co
Real estate, renting, busin
Health and social work
Other community, social ar
0
1
2
3
4
5
6
7
8
9

Data prepared for analysis

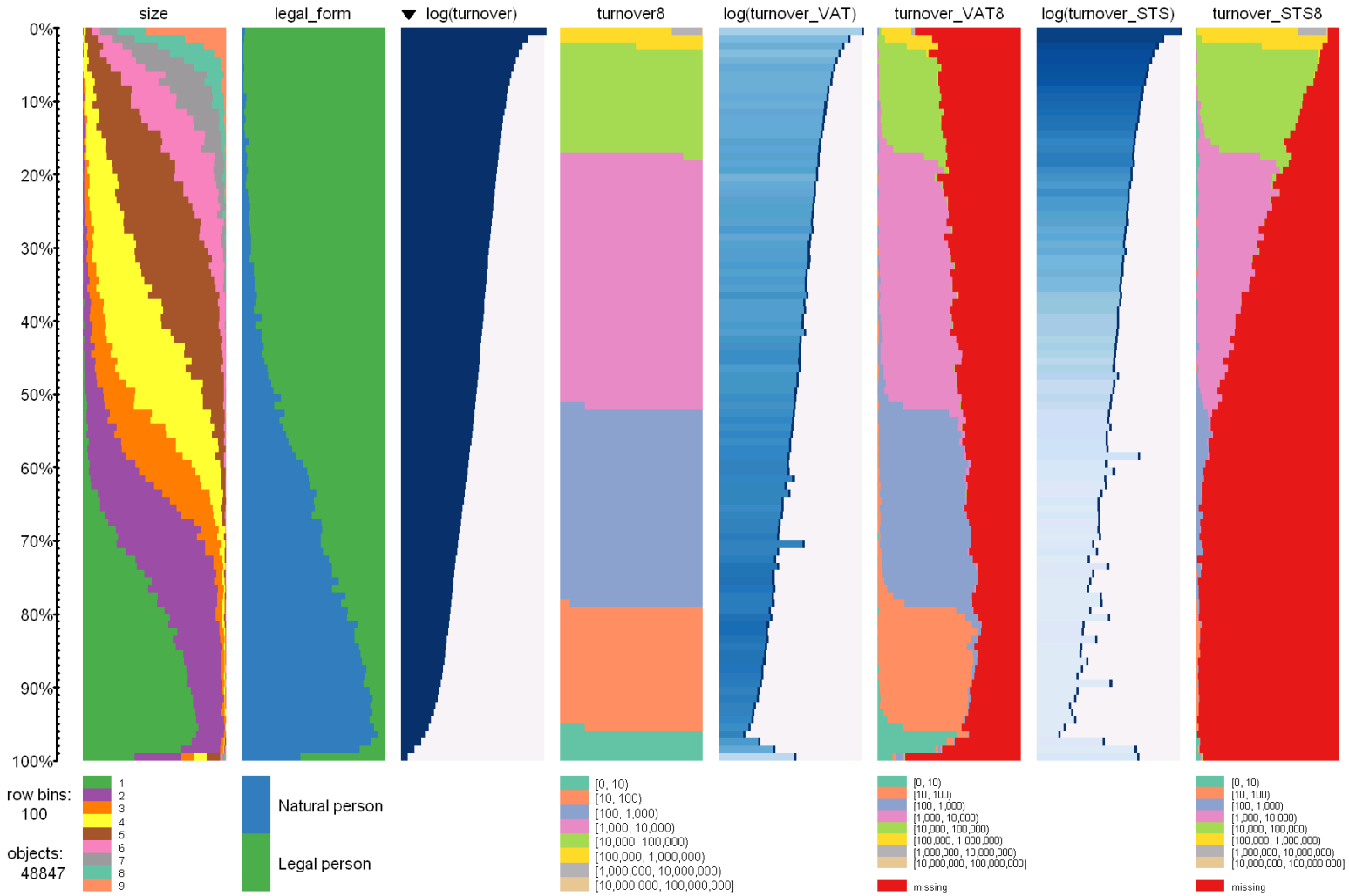


Comparison with other sources

Comparison:

- SBS turover
- VAT turnover
- STS turnover

Comparison with other sources



Implementation in R

- Package **tabplot**
- Available on CRAN
- Functions

tableplot	tableplot(myDataFrame, colNames = myColumnNames, nBins = 100)
num2fac	num2fac(myNumericVector, method="pretty", n=5)
tabGUI	tableGUI()

- Supports very large datasets (up to 2.10^9 records)

Conclusion

- Quality assessment
  - Existing data sources
  - New data sources
- Effective method to support top-down data analysis
- Apply tableplot to other sources
- Further improve tableplots

Conclusion

- Quality assessment
  - Existing data sources
  - New data sources
- Effective method to support top-down data analysis
- Apply tableplot to other sources
- Further improve tableplots

Conclusion

- Quality assessment
  - Existing data sources
  - New data sources
- Effective method to support top-down data analysis
- Apply tableplot to other sources
- Further improve tableplots

Conclusion

- Quality assessment
  - Existing data sources
  - New data sources
- Effective method to support top-down data analysis
- Apply tableplot to other sources
- Further improve tableplots